JOIN (Jurnal Online Informatika)
Volume 7 No. 1 | June 2022: 1-6
DOI: 10.15575/join.v7i1.736

p-ISSN: 2527-1682
e-ISSN: 2527-9165

Data Visualization of COVID-19 Vaccination Progress and Prediction Using Linear Regression

Hilal H. Nuha[1] and Ahmad Abo Absa[2]
[1]HUMIC Engineering, School of Computing, Telkom University, Indonesia
[2]University of Palestine, Palestine

Article Info

Article history:

Received May 11, 2021
Revised August 02, 2021
Accepted August 31, 2021
Published June 30, 2022

Keywords:

COVID19
Data Visualization
Prediction

ABSTRACT

This paper provides a data visualization and analysis of the COVID-19 vaccination program. Important information such as which countries have the highest vaccination rates and numbers. In addition to the types of vaccines used and used by countries in the world, an infographic on the geographic distribution of vaccine use is also shown. To model the obtained data, daily vaccination rates were modeled by linear regression in which five sample countries with different vaccination ranges were processed using data science approach, namely, linear regression. The modeling results show a gradient coefficient that represents an increase in vaccine rates. The prediction results showed that the highest rate of increase in daily vaccination was 1826126 additional vaccines per day.

Corresponding Author:

Hilal H. Nuha
School of Computing, Telkom University, Indonesia
Email: hilalnuha@telkomuniversity.ac.id

1. INTRODUCTION

The Covid 19 pandemic is a disaster in the health sector that has hit the whole globe by reaching over 2.7 Million of death toll [1]. With the advent of the Covid19 vaccines, estimating the vaccination rate is an important issue. Most of the vaccines developed by scientists often cause curiosity regarding the manufacturing process and the process of distributing these vaccines. This is due to the large number of researchers, practitioners, statisticians, and medical personnel who have tried to keep up with the spread of the virus in various countries in the world using various methods.

Several studies related to COVID-19 vaccination related to death rates and campaign efforts to the community. Zanettini et.al.[2] discussed the COVID-19 mortality rate and influenza vaccination in the United States. Doubts about the use of the COVID-19 vaccine in susceptible people were discussed by Rolland et al. [3]. The use of the flu vaccine for children and its relationship to COVID-19 was reported by Pratico et al. [4]. Arokiaraj [5] discusses the correlation between influenza vaccination and influenza COVID-19 symptoms. A machine learning approach to the analysis of vaccination scheduling and COVID-19 mortality rates was discussed by Yousef et al. [6]. From these studies, one thing that can be observed is that there is no analysis of the visualization of vaccination rates. This paper follows the visualization approach provided by [7]. which uses a public dataset. This paper wishes to present an analysis of the worldwide COVID-19 vaccination program using data taken from the Internet [8]. The data is processed to show some important information, such as countries that started vaccination for the first time and countries with the highest vaccinations. In addition, the types of vaccines offered and used by countries in the world. The prediction of the number of vaccines to be given is also presented using data science approach, i.e., linear regression. This paper is presented in the following arrangement. Section II discusses the methodology. The experimental results and conclusions are given in Sections III and IV, respectively.

2. METHOD

This section presents the methodology used in the experiment. The initial section of this chapter discusses data pre-processing followed by data analysis. This chapter ends with an explanation of the linear regression method for prediction.

e-ISSN 2527-1682
p-ISSN 2527-9165

2.1. Pre-Processing

The downloaded raw data is read and collected based on several new fields, namely `country`, `iso_code`, and the `vaccine`, which is the vaccine scheme used in a particular country [9]. At this stage, an important procedure is carried out, namely data cleaning, which is an important step for data analysis. The case that often occurs is the appearance of the `Not-A-Number(NaN)` value which can be resolved by changing the value to 0. In addition, empty rows marked with a value of 0 and also repeating columns can be resolved by deleting the column manually, directly, or by deleting unwanted rows.

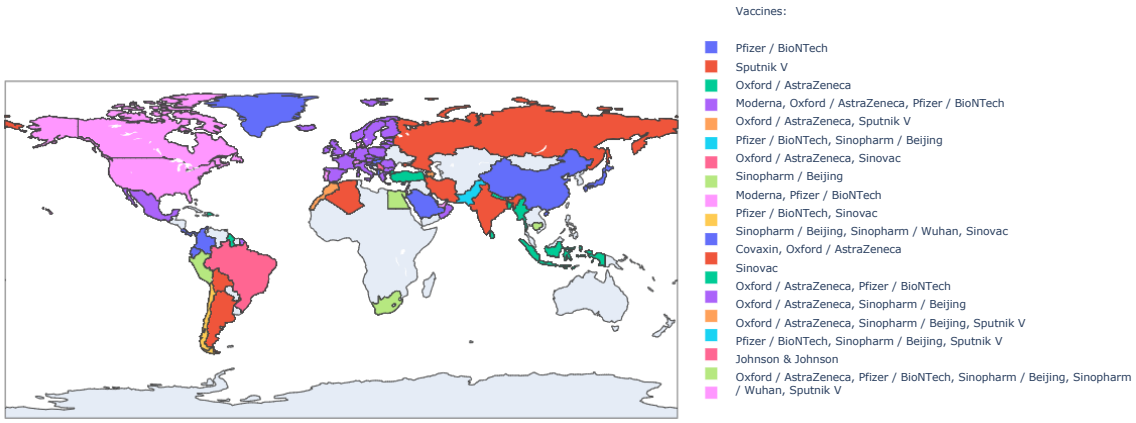

Figure 1. Vaccines used by various countries

2.2. The most widely used vaccine

The pre-processed and cleaned data is further sorted with respect to its country and vaccine types. Figure 1 shows the mapping of countries and types of vaccines used, for example, Moderna, Pfizer/BioNTech used by the USA, and Oxford / AstraZeneca used by the UK. In terms of the percentage of vaccines used worldwide, to obtain the types and in-depth understanding of vaccines used worldwide, visualization of vaccination data is carried out by descending sorting of vaccine use data, as shown in Figure 2.

Figure 3. shows the distribution and percentage of vaccine use used around the world. One can observe that the Moderna Pfizer vaccine has the largest portion of usage. The Sinovac vaccine is in the second position, where it can be easily inferred that the high usage is due to China's large population.

2.3. Countries with the highest vaccination rates

To display the vaccination data based on time, the values in the data are arranged in a certain order. The data compilation is done with the `groupby()` function to group data by country, `sort_values()` to sort the data based on the number of vaccinations, and the `max()` function to find the country with the highest number of vaccinations. Figure 4 shows the results for the ten countries with the highest vaccination percentages.

However, a developing country like Indonesia is able to surpass 10 million vaccine doses by achieving 500,000 doses per day. Since Indonesia is not yet a vaccine-producing country, the government is afraid of embargoes from other producing countries.

Figure 4 clearly shows that Israel is the only country with a percentage rate of vaccination per capita, with more than 80 people vaccinated for every 100 people. The UAE, UK, and USA are in second, third, and fourth positions, respectively. Israel and the UAE are countries with small populations; therefore, the vaccine administration is relatively easy to manage. The four countries have a percentage above 10%. Meanwhile, the fifth position is occupied by Turkey, which achieves a percentage of less than 10%. So it can be concluded that only four countries in the world vaccinate over 10%.

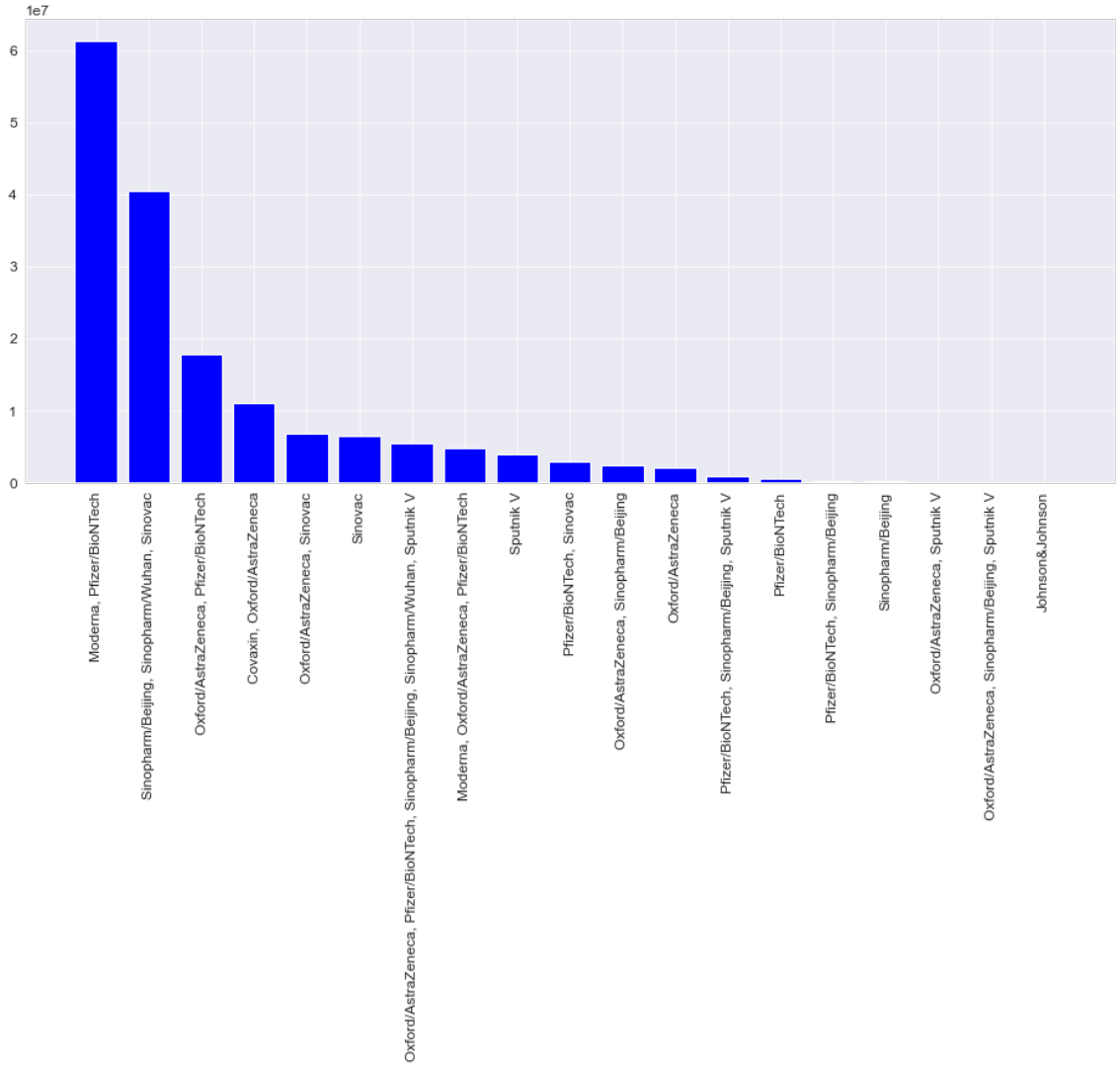

Figure 2. Most Used Vaccines

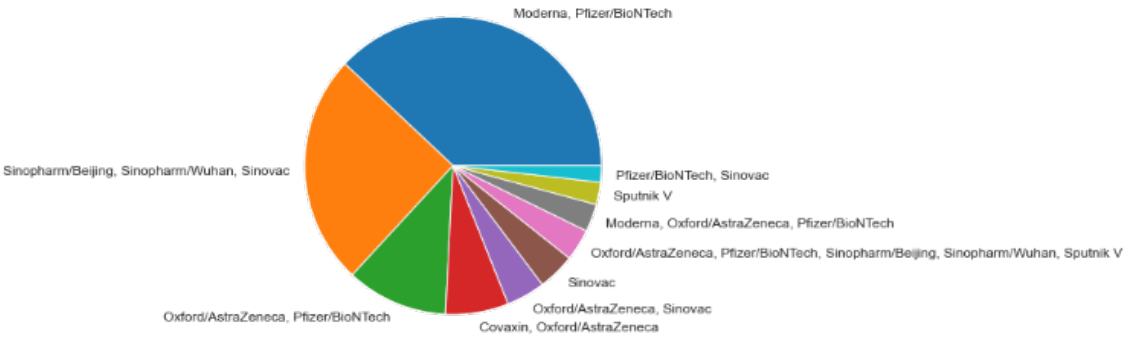

Figure 3. Percentage of 10 Most Frequently Used Vaccines

e-ISSN 2527-1682
p-ISSN 2527-9165

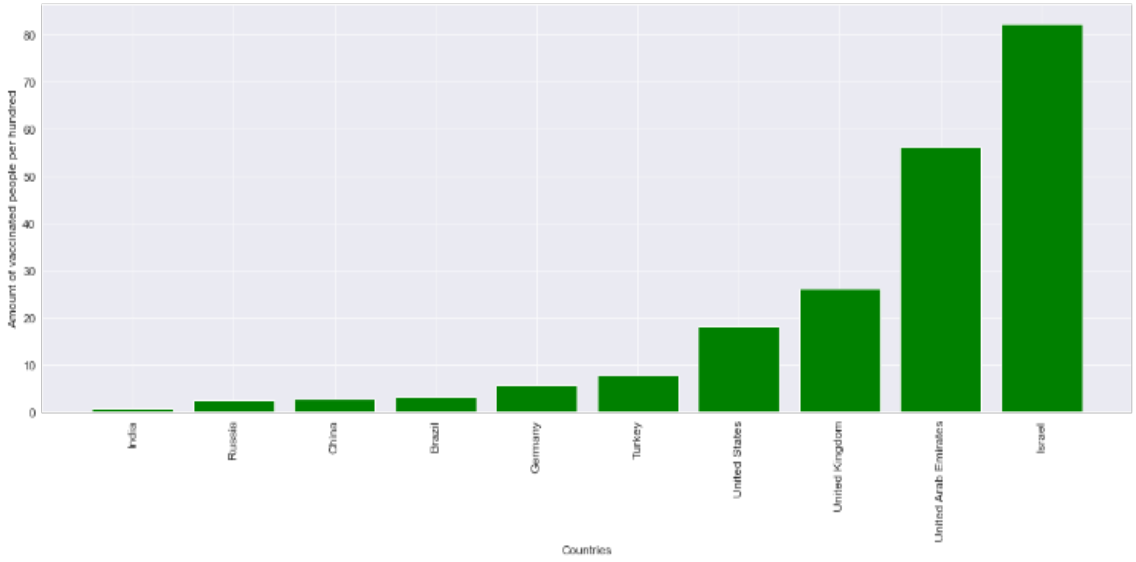

Figure 4. Ten Countries with the Highest Vaccination Percentage

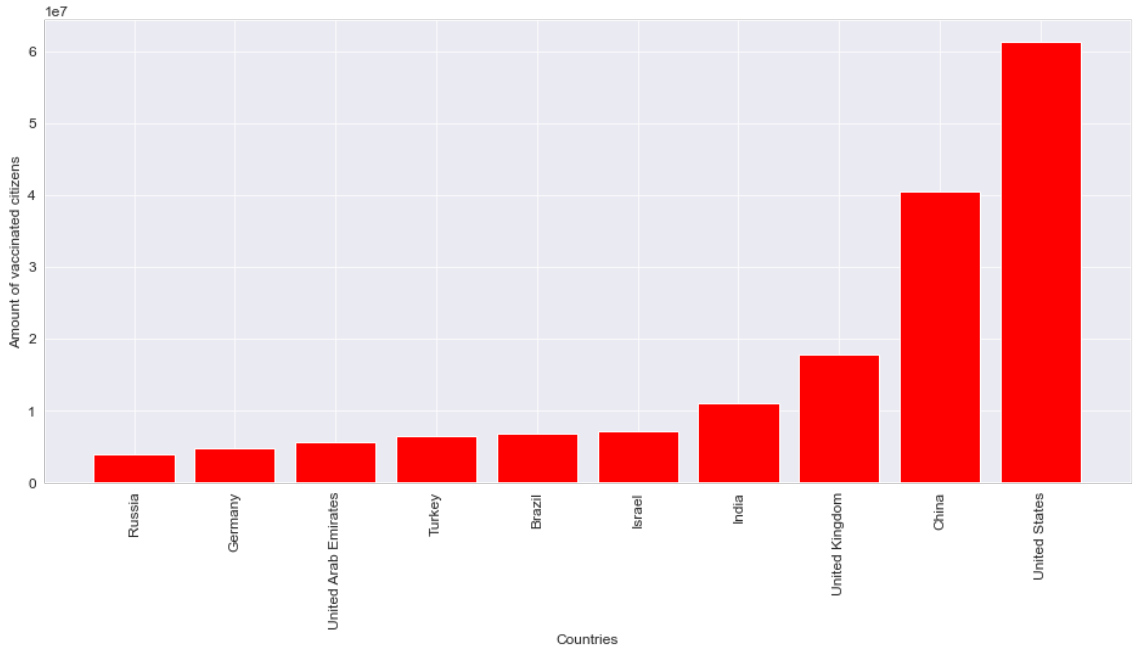

Figure 5. Total people vaccinated per country

2.4. Countries with the highest number of vaccinations

This section shows the ten countries with the highest number of vaccinated people. The matplotlib library is used to visualize data based on the total vaccinations.

Figure 5 shows that the US has a population of about 60 million people vaccinated. Developed countries such as China and UK are in second and third positions, respectively. These three countries are developed countries, so citizens have broad access to vaccines. Meanwhile, in developing countries such as Indonesia, the percentage of vaccination is still low. Figure 6. This figure shows the trend of daily vaccination in Indonesia.

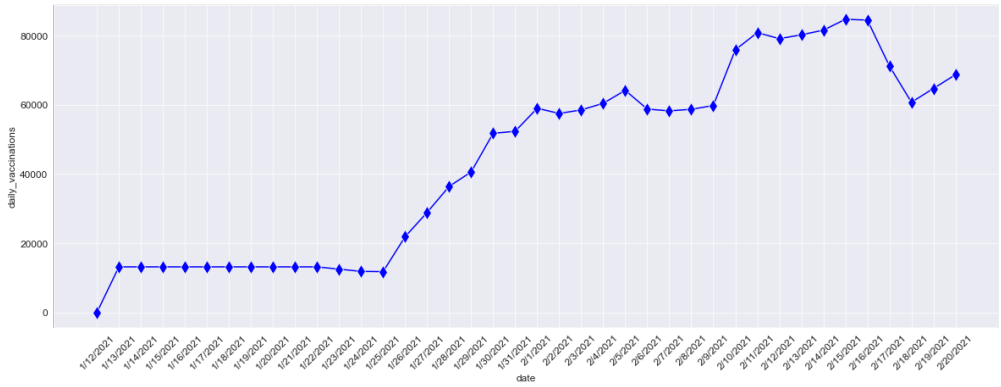

Figure 6. Indonesia's daily vaccinations population trend

2.5. Prediction of Number of Vaccinations Using Linear Regression

This sub-section presents an implementation of data science approach for vaccination. The vaccination rate per day is presented using a linear function of time i.e., day. Simple linear regression techniques can be used to predict the number of vaccinations. Linear prediction coding (LPC) [10], [11] differs from linear regression by the use of single dimensionality. LPC is usually applied to one dimension time series without taking other values as inputs. Meanwhile, linear regression assumes inputs as linear combinations. Simple linear regression models data as a function as a linear curve or an affine function.

$$y_i = ax_i + b + c_i \quad (1)$$

where is the output value in this case the vaccination rate and is the number of days from the first time the vaccine was administered. The coefficients such as y_i and x_i are the slope coefficient and constant, respectively. The error or difference from each point is given by c_i.

Linear regression tries to find the values of a and b so that the cost function is the total value of the error squared E according to the following equation:

$$E = \sum_{i=1}^{N} c_i^2 \quad (2)$$

where N is the number of samples. The error value c_i can be written as follows:

$$c_i = y_i - (ax_i + b) \quad (3)$$

By using the partial derivative of the gradient $(\frac{\partial E}{\partial a})$, the value of slope a gets the smallest value of E, i.e.

$$a = \frac{\sum_{n=1}^{N} \left(x_n - \underline{x}\right)\left(y_n - \underline{y}\right)}{\sum_{n=1}^{N} \left(x_n - \underline{x}\right)^2} \quad (4)$$

and the value of the constant b which minimizes E is also obtained by using a partial derivative $\left(\frac{\partial E}{\partial b}\right)$of the constant so that the optimum value of b is

$$b = \underline{y} - a\underline{x} \quad (5)$$

where $\underline{x}$ is the average value of the inputs

$$\underline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad (6)$$

and $\underline{y}$ is the average value of the outputs

$$\underline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n. \quad (7)$$

3. RESULTS AND DISCUSSION

To make predictions against a linear model, data are retrieved from [8] to be used in the linear model. The data shows daily vaccination rates in different countries up to the same date. Because each country starts vaccination on a different date but is measured until the same date, the number of data (N) has a different value for each country. In this paper, five countries were selected as samples for prediction, namely Indonesia, India, USA, United Kingdom, and China.

e-ISSN 2527-1682
p-ISSN 2527-9165

Table 1. Prediction Result

no	Country	N	a	b	Predicted $y(N+1)$
1	Indonesia	38	2246.29	903.29	88508.56
2	India	36	7110.89	149130.39	412233.39
3	USA	62	**29007.35**	-1335.76	**1826126.95**
4	Great Britain	**68**	7381.47	-7736.85	501585.15
5	China	56	26013.91	-117780.44	1365012.44

Table I. shows the details of the data, the modeling coefficient of linear regression results, and the prediction results. Because data collection was stopped on the same day, it shows which country was the earliest to vaccinate. The table shows that the country with the largest N is Great Britain with 68 days of vaccine administration. The slope coefficient a represents daily increase so that the country with the highest daily increase, namely the United States, is also predicted to have the highest value at $N + 1$. Great Britain has a relatively smaller population size than that of the other sample countries; therefore, approximately 7381 daily doses of vaccine are sufficient. Compared to Indonesia, India starts the vaccination a little bit later. However, India's vaccination rate is significantly higher than Indonesia's. Therefore, the predicted vaccination rate of India is almost five times of Indonesia.

4. CONCLUSION

Data visualization and analysis of the COVID-19 vaccination program are provided. Important information such as which countries achieve the highest vaccination rates and numbers. In addition to the types of vaccines used and used by countries in the world, an infographic on the geographic distribution of vaccine use is also shown. To model the data held, daily vaccination rates were modeled by linear regression in which five sample countries with different vaccination ranges were processed using linear regression. The modeling results show a gradient coefficient that represents an increase in vaccine rates. The prediction results showed that the highest rate of increase in daily vaccination was 1826126 additional vaccines per day achieved by the United States.

5. REFERENCES

[1] R. Lingga and H. H. Nuha, "Accessibility and Response Time Analysis on the COVID19 Website in Indonesia," 2021.
[2] C. Zanettini *et al.*, "Influenza vaccination and COVID19 mortality in the USA," *medRxiv*. 2020, doi: 10.1101/2020.06.24.20129817.
[3] Y. Rolland, M. Cesari, J. E. Morley, R. Merchant, and B. Vellas, "COVID19 Vaccination in Frail People. Lots of Hope and Some Questions," *Journal of Nutrition, Health and Aging*. 2021, doi: 10.1007/s12603-021-1591-9.
[4] A. D. Praticò and M. Ruggieri, "In Covid19-era, is it time for a mandatory Flu-vaccination for children?," *Science (80-.).*, 2020.
[5] M. C. Arokiaraj, "Correlation of Influenza Vaccination and the COVID-19 Severity," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3572814.
[6] M. Yousef, L. Showe, and I. Ben Shlomo, "Machine Learning Analysis of National Vaccination Schedules and Rates of Compliance Reveals Correlation with COVID19 Mortality," *Int. J. Comput. Biol.*, vol. 9, pp. 1–6, 2020.
[7] N. Sutaria, "COVID-19 Vaccination Progress Analysis through Data Storytelling," 2021. https://towardsdatascience.com/covid-19-vaccination-progress-analysis-around-the-world-736d7e57f198 (accessed Mar. 19, 2021).
[8] "COVID-19 World Vaccination Progress." https://www.kaggle.com/gpreda/covid-world-vaccination-progress/ (accessed Mar. 29, 2021).
[9] G. Press, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," *Forbes Tech*, 2016.
[10] B. Liu *et al.*, "A multi tone modeling for seismic data compression," in *SEG Technical Program Expanded Abstracts*, 2019, pp. 263–267.
[11] B. Liu, M. Mohandes, H. Nuha, M. Deriche, F. Fekri, and J. H. McClellan, "A Multitone Model-Based Seismic Data Compression," *IEEE Trans. Syst. Man, Cybern. Syst.*, 2021, doi: 10.1109/tsmc.2021.3077490.